Integration with KMS for Controlled Vocabulary

Goal

In order to support hierarchical facets for collections we will need to utilize the full hierarchy for providers, science keywords, platforms, and instruments. For science keywords we receive the full hierarchy as part of the collection metadata so it is not necessary to use the GCMD Keyword Management System (KMS). However we do not receive the full hierarchy as part of the metadata for providers, platforms, or instruments. That is the driver for integrating with KMS over the short term.

In the long term we will need to begin validating metadata records against the KMS controlled vocabulary.

Traceability

⚠ *CMR-1830 - JIRA project doesn't exist or you don't have permission to view it.*

Hierarchical Facets Use Case (Near term)

GCMD provides a drill-down capability to find collections the user is interested in by utilizing facets. See http://gcmd.nasa.gov/KeywordSearch/Keywords.do?Portal=GCMD&MetadataType=0&Columns=0&KeywordPath=DataCenters. In order to continue to provide this functionality while hosted on top of the CMR, the CMR needs to return hierarchical facets for all of the fields mentioned above.

The end result is that curl https://cmr.sit.earthdata.nasa.gov/search/collections.json?include_facets=true&hierarchical_facets=true needs to return facets in a way that supports the drill-down capability.

In order to support this use case several CMR components are affected. Note that I use providers/archive centers in the examples below. Support will be added the same way for platforms and instruments.

Indexer

- When indexing a collection retrieve the hierarchy based on the archive center short name provided in the metadata.

  > This requires the short names to be guaranteed to be unique in the KMS controlled vocabulary. GCMD has agreed this will be the case and will be enforced in the KMS in the future.

- Index the full platform, archive center, and instrument hierarchies (similar to science keywords also index all sub-fields as lowercase)

Search

- Need to produce hierarchical facets
- Need to support searching by subfields

```
• {"condition" : {"archive_center": {"level_1": "NASA"}}}
```

General

Retrieve Full Hierarchy from KMS

- Use the static CSV file on the GCMD website that contains the full hierarchy for providers. The file is updated 4 times a day (at 0:00, 6:00, 12:00, and 18:00).
- Use a background job to refresh the controlled vocabulary every 2 hours. This ensures that the keywords used in CMR are never more than 8 hours old.
- Save in memory as a nested map with the short-name as the key for each entry.

```
{"LP DAAC" {:uuid "5612b95e-cc41-4286-9946-e78506f70f59",
        :data-center-url "https://lpdaac.usgs.gov/",
        :long-name "Land Processes Distributed Active Archive Center",
        :short-name "LP DAAC",
         :level-1 "NASA",
        :level-0 "GOVERNMENT AGENCIES-U.S. FEDERAL AGENCIES"}
 "NSIDC" {...}
 ...}
```

Caching

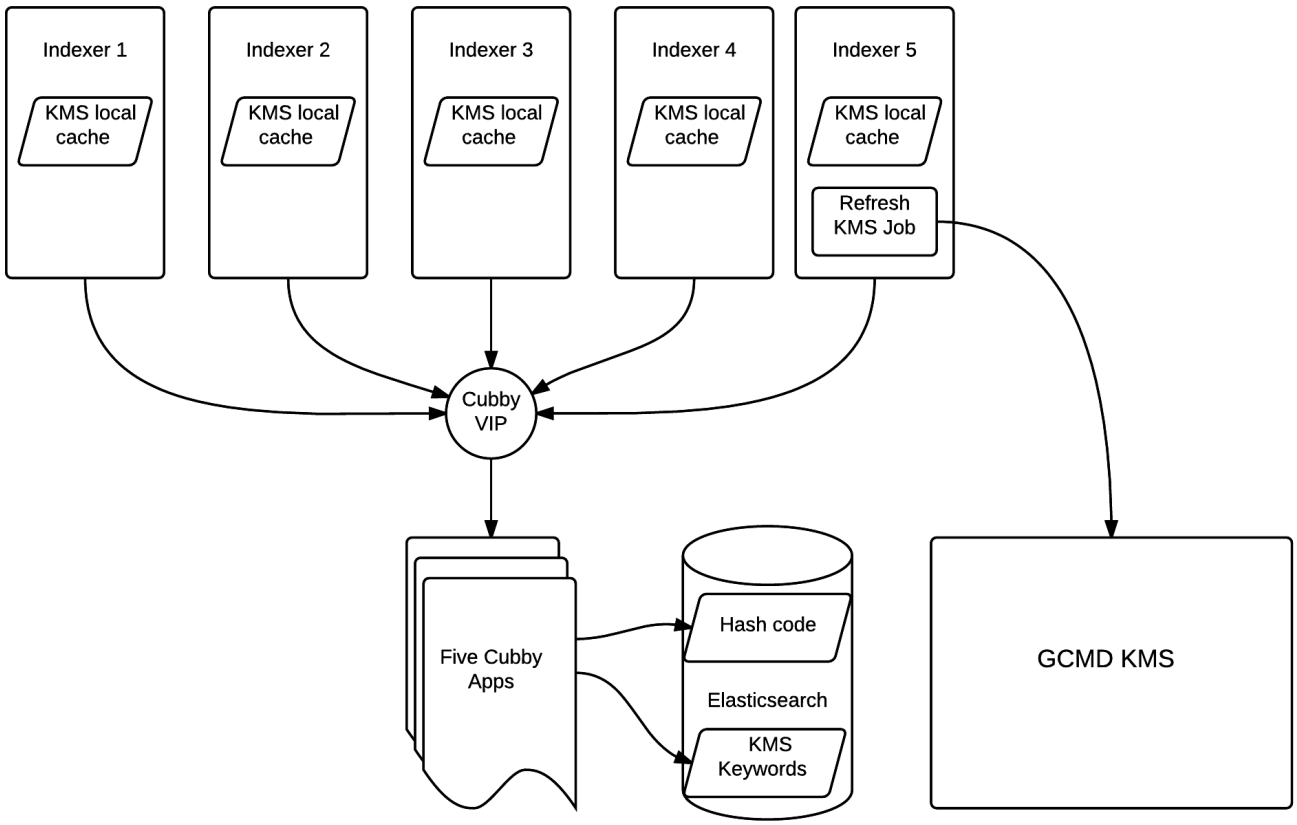
We cache the provider keywords to avoid retrieving them from GCMD on every indexing request.

Considerations

1. In the event of downtime on the GCMD website we need to be able to start up successfully and be able to index collections.
2. We have multiple indexing nodes.
3. When there is an update to the keywords we want all nodes to use the new keywords on all indexing nodes.
4. In the future ingest will also need to cache the keywords in order to perform validation of the metadata against the keywords.

Design

We use a strategy similar to the way we cache ACLs. We cache the keywords locally on each application node and use Cubby to determine if there have been any changes to the KMS keywords. We do this by tracking the hash code of the full keywords list with Cubby. If the hash code for the keywords in cache does not match the hash code in Cubby we will retrieve the latest keywords. The main difference from the way ACLs work is that instead of retrieving the keywords from GCMD we will instead retrieve the keywords directly from Cubby (if Cubby does not have the keywords cached we will go to the GCMD site). So Cubby will have both the hash code as well as the full keyword hierarchy indexed. When the background job refreshes the keywords from the GCMD site, it will save the updated hierarchy to Cubby in addition to saving the updated hash code.



Local Testing

Rather than retrieving the keywords from GCMD we will instead save local copies and make them available via HTTP within dev-system. Both local development and CI will utilize these copies.

Open issues

1. Archive center names do not match provider short names in KMS and we are not guaranteed that all platform and instrument short names are in KMS. This will be addressed as part of reconciliation, GCMD will also be able to mark data as curated or not when figuring out what to display on their site and portals. We opened CMR-1944 to still return platforms in the hierarchical facets which do not have entries in KMS under a dummy hierarchy.
2. There are 6 duplicate instrument short names, 1 duplicate provider short name, and 3 duplicate platform short names in KMS. There are also some entries which do not match the KMS definition for that type. The GCMD team will address.

Error rendering macro 'pageapproval' : null